

## CLAIMS

What is claimed is:

- 1 1. A method comprising:  
2 representing an input document image with a sequence of template identifiers to  
3 reduce storage consumed by the input document image; and  
4 replacing the template identifiers with alphabet characters according to language  
5 statistics to generate a text string representative of text in the input document  
6 image.
- 1 2. The method of claim 1 further comprising extracting n-gram indexing terms from  
2 the text string by selecting alphabet characters in the text string that satisfy a  
3 predicate and then combining the alphabet characters to form n-grams, n being an  
4 integer.
- 1 3. The method of claim 2 wherein the n-grams are trigrams.
- 1 4. The method of claim 2 wherein the number of characters in the text string that  
2 satisfy the predicate is fewer than the number of characters in the text string.
- 1 5. The method of claim 2 wherein the predicate is a condition that substantially all  
2 selected characters follow respective spaces in the text string.
- 1 6. The method of claim 2 further comprising comparing the input document image  
2 with a plurality of symbolically compressed document images in a database based

3 on the n-gram indexing terms to determine whether one of the plurality of  
4 documents matches the input document image.

1 7. The method of claim 6 further comprising determining whether the one of the  
2 plurality of document images is a confidential document if the one of the plurality  
3 of document images matches the input document image.

1 8. The method of claim 7 further comprising prompting a user for authorization  
2 before operating on the input document image if the one of the plurality of  
3 document images matches the input document image and is a confidential  
4 document.

1 9. The method of claim 6 further comprising determining whether the one of the  
2 plurality of document images forms a sub-portion of an encompassing document  
3 image if the one of the plurality of document images matches the input document  
4 image.

1 10. The method of claim 6 further comprising prompting the user to select between  
2 hardcopy output of the one of the plurality of document images or the  
3 encompassing document image if the one of the plurality of document images  
4 matches the input document image and forms a sub-portion of a larger document  
5 image.

1 11. The method of claim 6 further comprising determining whether the one of the  
2 plurality of document images is copyrighted if the one of the plurality of  
3 document images matches the input document image.

- 1 12. The method of claim 1 wherein replacing the template identifiers with alphabet  
2 characters comprises replacing at least one of the template identifiers with an  
3 alphabet character selected according to a sequence of at least two alphabet  
4 characters selected to replace template identifiers that precede the at least one  
5 template identifier in the sequence of template identifiers.
- 1 13. The method of claim 1 wherein replacing the template identifiers with alphabet  
2 characters comprises replacing at least one of the template identifiers with an  
3 alphabet character selected according to a sequence of all alphabet characters  
4 selected to replace template identifiers that precede the at least one template  
5 identifier in the sequence of template identifiers.
- 1 14. The method of claim 1 wherein replacing the template identifiers with alphabet  
2 characters according to language statistics comprises replacing the template  
3 identifiers with alphabet characters selected according to a hidden Markov model.
- 1 15. The method of claim 1 wherein replacing template identifiers with alphabet  
2 characters according to language statistics comprises solving a substitution cipher  
3 by mapping the alphabet characters to the template identifiers based at least partly  
4 on frequency of occurrence of the template identifiers.
- 1 16. A method comprising using a hidden Markov model to solve a substitution cipher.
- 1 17. A document processing system comprising:  
2 a deciphering module to generate a first text string based on a sequence of  
3 template identifiers in a first symbolically compressed document image and

4 to generate a second text string based on a sequence of template identifiers in  
5 a second symbolically compressed document image;  
6 a conditional n-gram module coupled to receive the first and second text strings  
7 from the deciphering module, the conditional n-gram module being  
8 configured to extract n-gram indexing terms from the first and second text  
9 strings based on a predicate condition; and  
10 a comparison module to generate a measure of similarity between the first and the  
11 second symbolically compressed document image based on the indexing  
12 terms extracted by the conditional n-gram module.

1 18. The document processing system of claim 17 wherein at least one of the  
2 deciphering module and the conditional n-gram module is implemented by a  
3 programmed processor.

1 19. The document processing system of claim 17 wherein the deciphering module  
2 generates the first text string by applying a hidden Markov model to the sequence  
3 of template identifiers in the first symbolically compressed document image.

1 20. The document processing system of claim 17 wherein the second symbolically  
2 compressed document image is obtained from a database of symbolically  
3 compressed document images.

1 21. The document processing system of claim 17 wherein the data processing system  
2 further comprises a scanning and compressing module that is configured to  
3 generate a digitized version of a source document and to perform symbolic  
4 compression of the digitized version to produce the first symbolically compressed  
5 document image.

- 1 22. The document processing system of claim 21 wherein the document processing  
2 system is a document copying system.
- 1 23. The document processing system of claim 21 wherein the document processing  
2 system is a facsimile transmission system.
- 1 24. The document processing system of claim 17 further comprising a second interface  
2 to couple the document processing system to a database of symbolically  
3 compressed document images and associated indexing terms, and wherein the  
4 conditional n-gram module is configured to store the n-gram indexing terms  
5 extracted from the second text string in the database.
- 1 25. The document processing system of claim 24 wherein the comparison module is  
2 configured to receive the n-gram indexing terms extracted from the first text string  
3 from the conditional n-gram module and to receive the n-gram indexing terms  
4 extracted from the second text string from the database via the second interface.
- 1 26. The document processing system of claim 17 wherein the measure of similarity is  
2 used to determine whether the first and second symbolically compressed  
3 document images match.
- 1 27. The document processing system of claim 26 wherein the data processing system  
2 further comprises a user interface to prompt a user to select between a  
3 decompressed version of the first symbolically compressed document image and a  
4 decompressed version of a third symbolically compressed document image that  
5 encompasses the second symbolically compressed document image if the first and

6 second symbolically compressed document images match, the document  
7 processing system further comprising an output module to output the  
8 decompressed version of the first symbolically compressed document image or the  
9 decompressed version of a third symbolically compressed document image  
10 according to input received via the user interface.

1 28. The document processing system of claim 26 wherein the data processing system  
2 further comprises a security module to determine if the second symbolically  
3 compressed document image is a confidential document based on attribute  
4 information associated with the second symbolically compressed document image,  
5 the security module being configured to prompt a user to enter authorization  
6 information before permitting output of the first symbolically compressed  
7 document image if the second document image is a confidential document and if  
8 the first and second document images match.

1 29. The document processing system of claim 28 further comprising a printer  
2 configured to receive a signal from the security module indicating whether to print  
3 a decompressed version of the first symbolically compressed document image.

1 30. The document processing system of claim 28 further comprising an transmission  
2 module to receive a signal from the security module indicating whether to  
3 transmit the first symbolically compressed document image.

1 31. The document processing system of claim 26 wherein the data processing system  
2 further comprises an monitoring module to determine if the second symbolically  
3 compressed document image is a copyrighted document.

1 32. The document processing system of claim 31 wherein the monitoring module is  
2 configured to automatically charge a copyright license fee for output of a  
3 decompressed version of the first symbolically document image if the second  
4 symbolically compressed document image is a copyrighted document and if the  
5 first and second symbolically compressed document images match.

1 33. An apparatus comprising a deciphering module to apply a hidden Markov model  
2 to decipher a sequence of template identifiers in a symbolically compressed  
3 document to recover a text string from the symbolically compressed document.

1 34. A method of extracting n-grams from a text, the method comprising:  
2 automatically selecting alphabetic characters in the text that satisfy a predicate;  
3 and  
4 concatenating the selected alphabetic characters to form n-grams.

1 35. The method of claim 35 wherein the predicate is a condition that the selected  
2 alphabetic characters follow respective spaces in the text.

1 36. An article of manufacture including one or more computer-readable media that  
2 embody a program of instructions to generate a text string from an input  
3 document image represented by a sequence of template identifiers for the purpose  
4 of reducing storage consumed by the input document image, wherein the program  
5 of instructions, when executed by one or more processors in the processing  
6 system, causes the one or more processors to replace the template identifiers with  
7 alphabet characters according to language statistics to generate a text string  
8 representative of text in the input document image.

1 37. The article of claim 36 wherein the one or more computer-readable media include  
2 one or more non-volatile storage devices

1 38. The article of claim 36 wherein the one or more computer-readable media include  
2 a propagated data signal.

1 39. The article of claim 36 wherein the program of instructions, when executed by the  
2 one or more processors in the processing system, causes the one or more  
3 processors to extract n-gram indexing terms from the text string by selecting  
4 alphabet characters in the text string that satisfy a predicate and then combining  
5 the alphabet characters in n-grams, n being an integer.

1 40. An article of manufacture including one or more computer-readable media that  
2 embody a program of instructions to generate a text string from an input  
3 document image represented by a sequence of template identifiers for the purpose  
4 of reducing storage consumed by the input document image, wherein the program  
5 of instructions, when executed by one or more processors in the processing  
6 system, causes the one or more processors to using a hidden Markov model to  
7 solve a substitution cipher formed by the sequence of template identifiers.